Saturdays.AI

Saturdays.AI
Donostia

#3 Decision Trees & Random Forests

by Saturdays.AI

Saturdays.AI
Machine Learning

Saturdays.AI
Donostia

Schedule

State of the course

Session 3 Review

Challenge

Notebook + resources

State of the course

Saturdays.AI Donostia

#1 Cleaning & Exploratory Data Analysis ✔️

#2 Supervised Learning ✔️

#3 Decision Trees & Random Forest 🔴 Today!

#4 Unsupervised Learning + Clustering 🔜

#5 Time Series Analysis + Data Viz 🔜

#6 Neural Networks, Gradient Descent 🔜

#7 NLP 🔜

Questions?

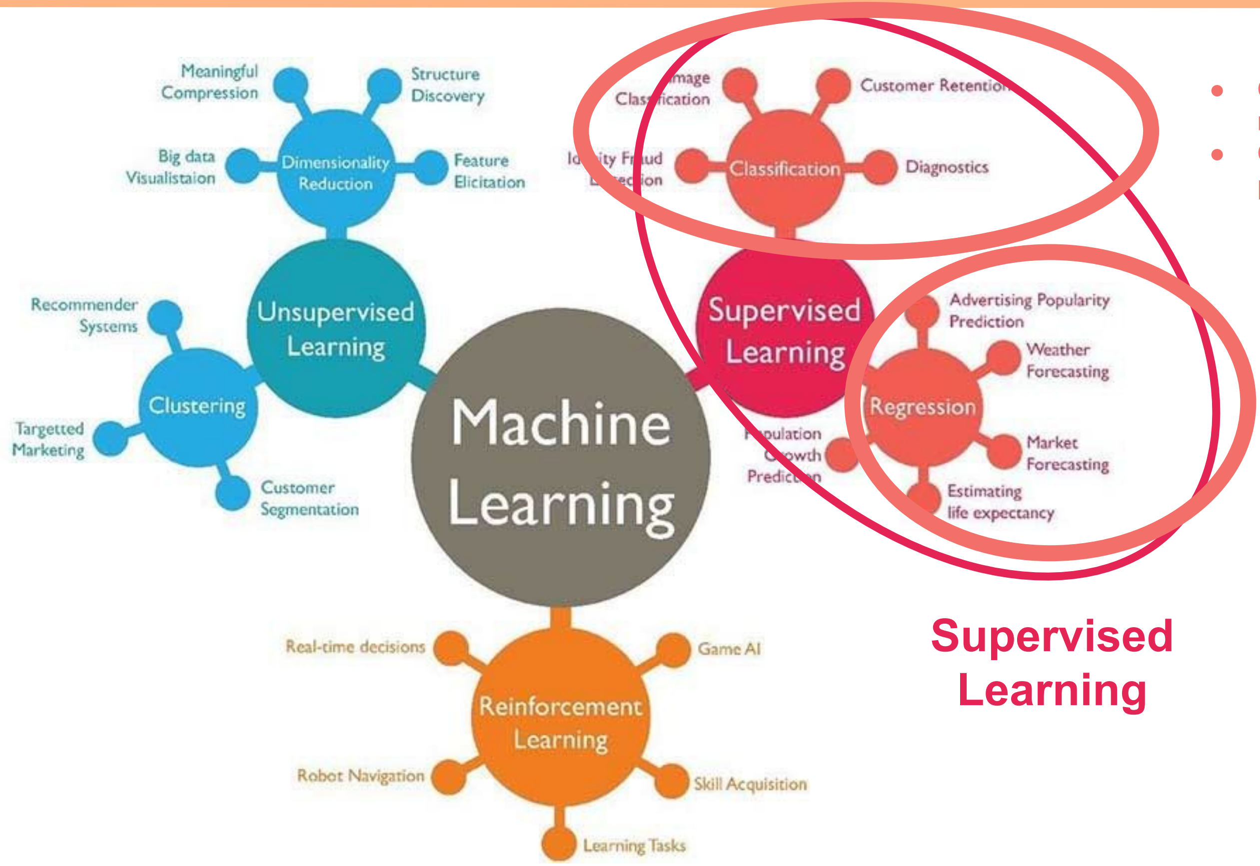
Saturdays.AI
Donostia

Decision Trees & Random Forest Deep Dive 🏊

Saturdays.AI
Donostia

- Decision Trees
- Random Forest
- Bagging, Boosting & Out of Bag
- Gradient Boosting
- Hyperparameter Tuning
- Feature Engineering
- Classification vs Regression Evaluation

¿Where we are?

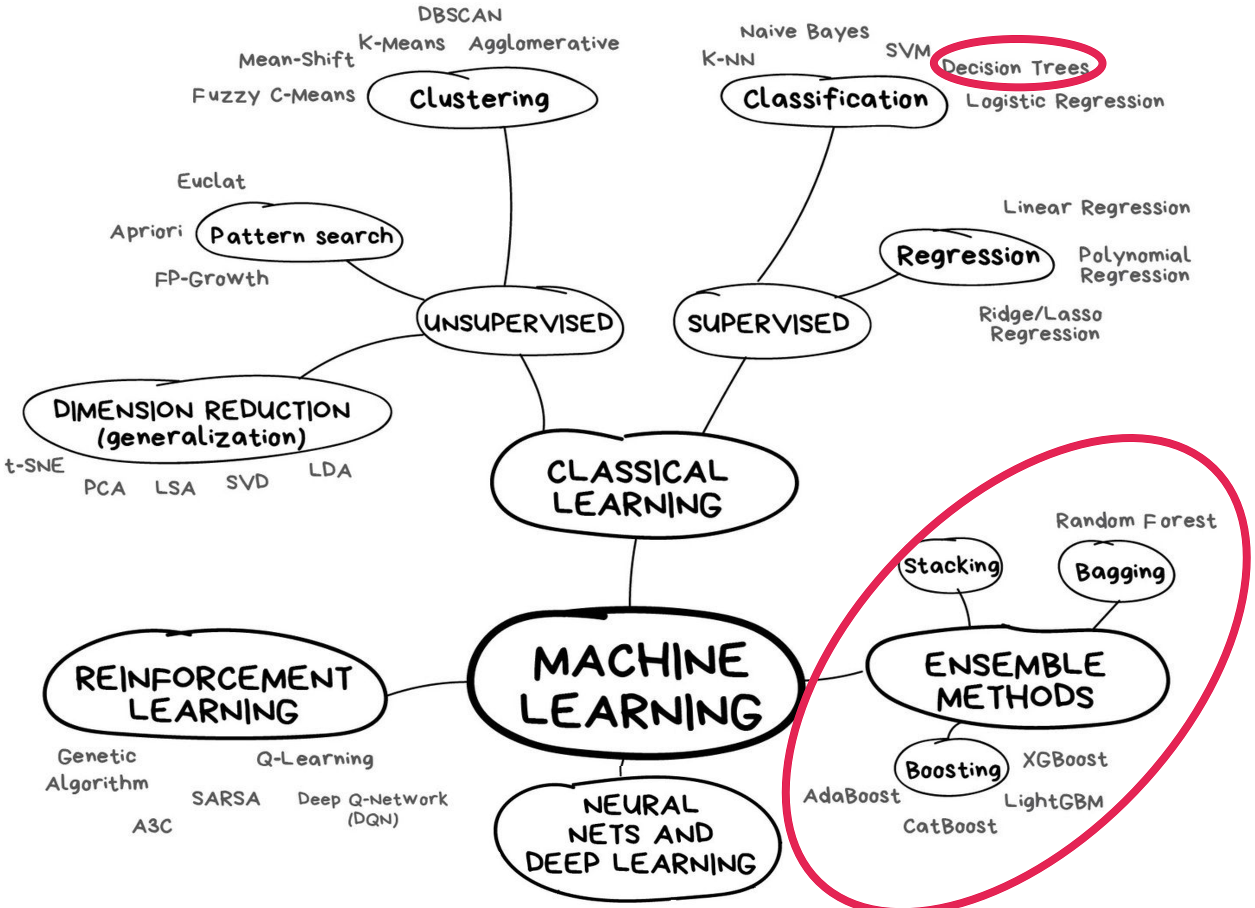


- Categorical response variable
- Classification problems

- Linear response variables
- Linear problems

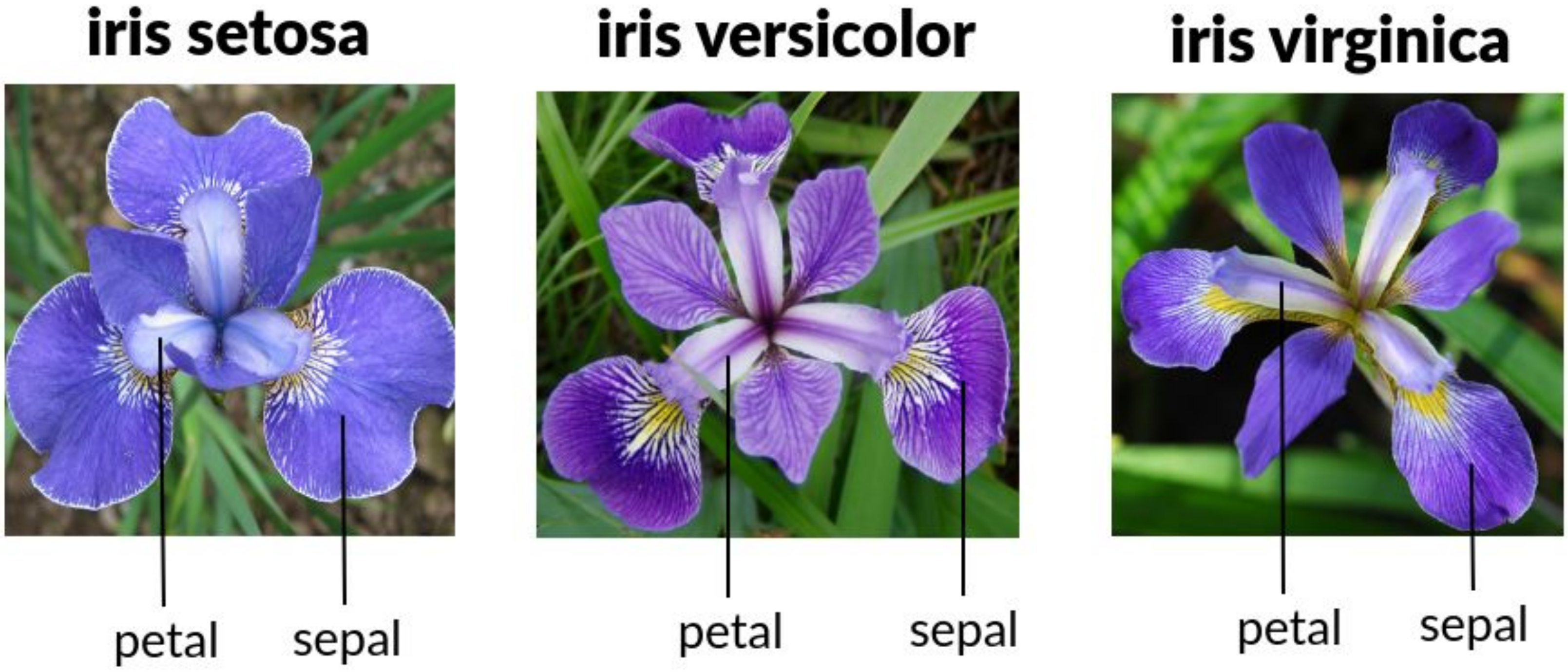
Supervised Learning

ML Algorithms

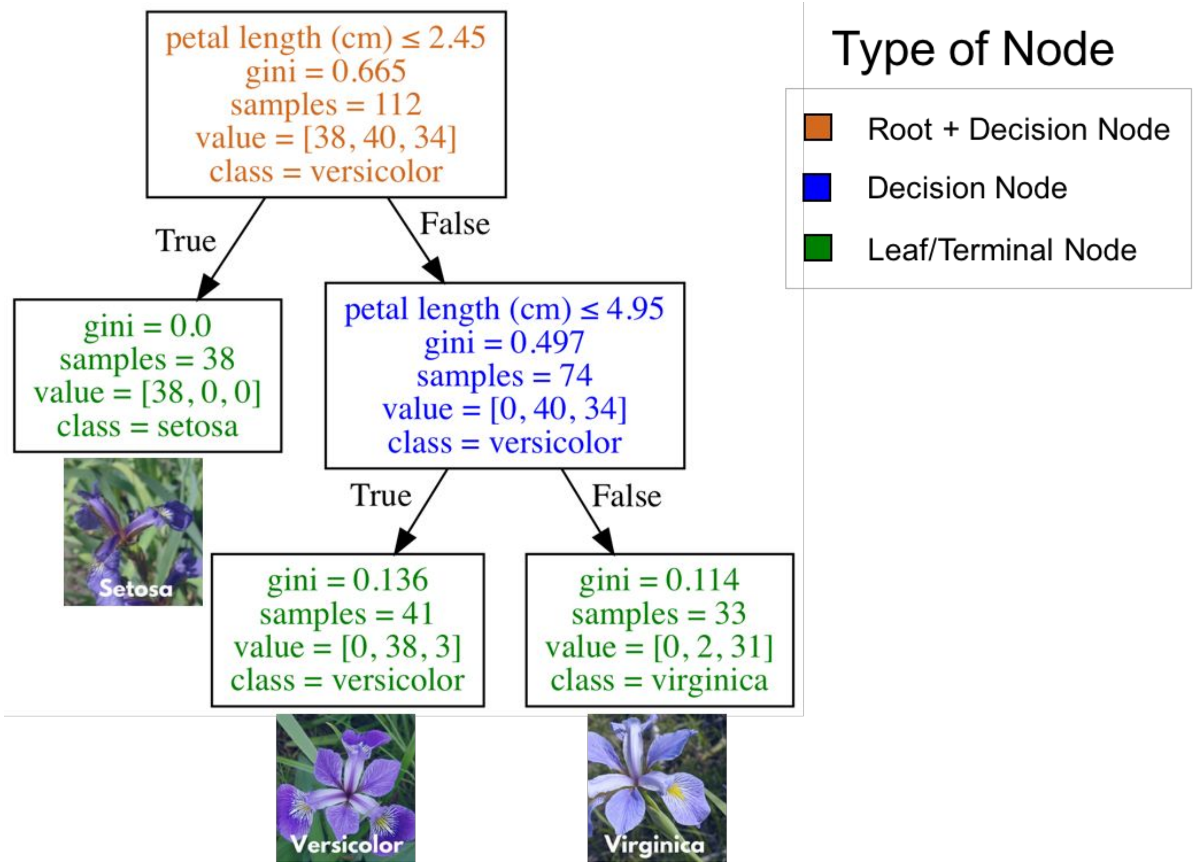


Saturdays.AI
Donostia

Decision Tree Algorithms



Decision Tree Algorithms



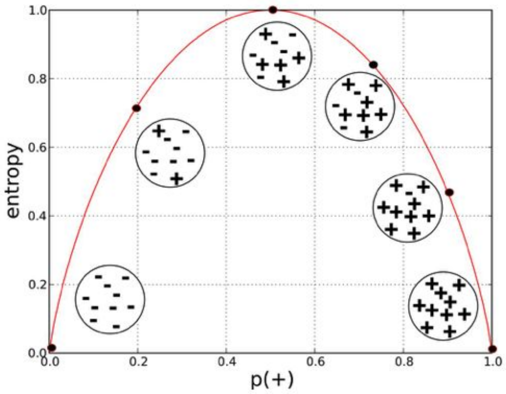
Decision Tree Algorithms: Concepts

GINI IMPURITY:

The **Gini impurity** can be computed by summing the probability of an item with label being chosen times the probability of a mistake in categorizing that item. It reaches its minimum (zero) when all cases in the node fall into a single target category.

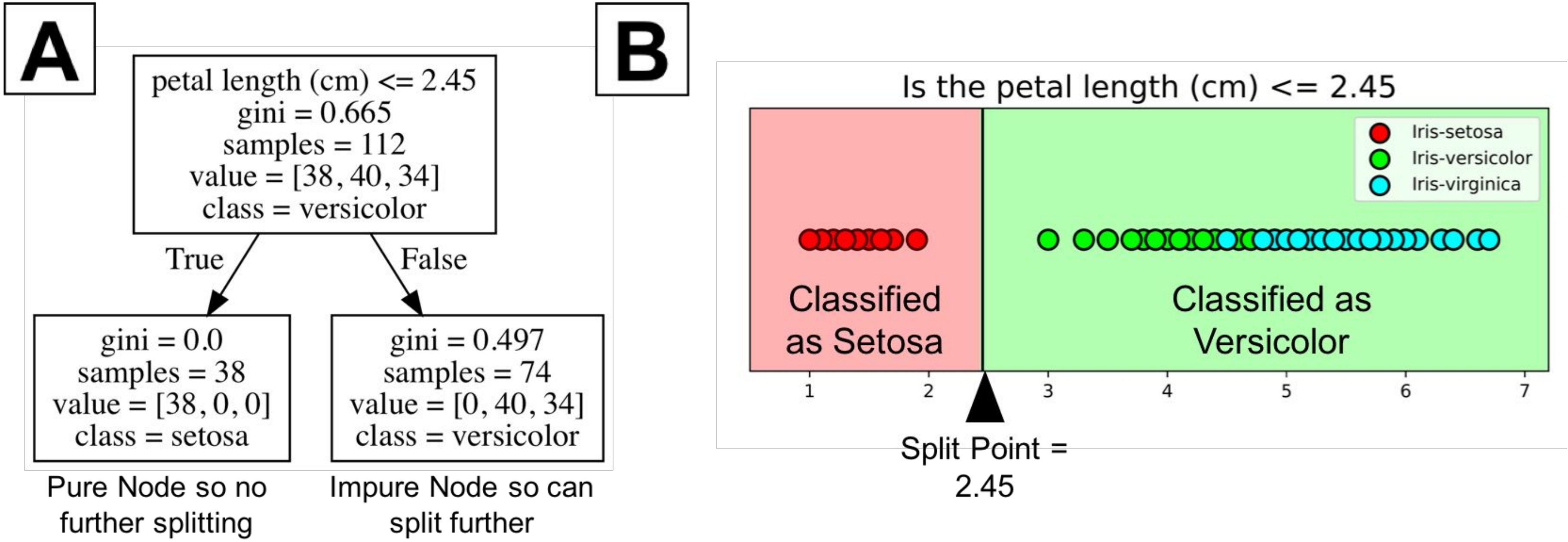
INFORMATION GAIN + ENTROPY:

Entropy to calculate the homogeneity (or impurity) of a sample. If the sample is completely homogeneous the entropy is zero and if the sample is an equally divided it has entropy of one.

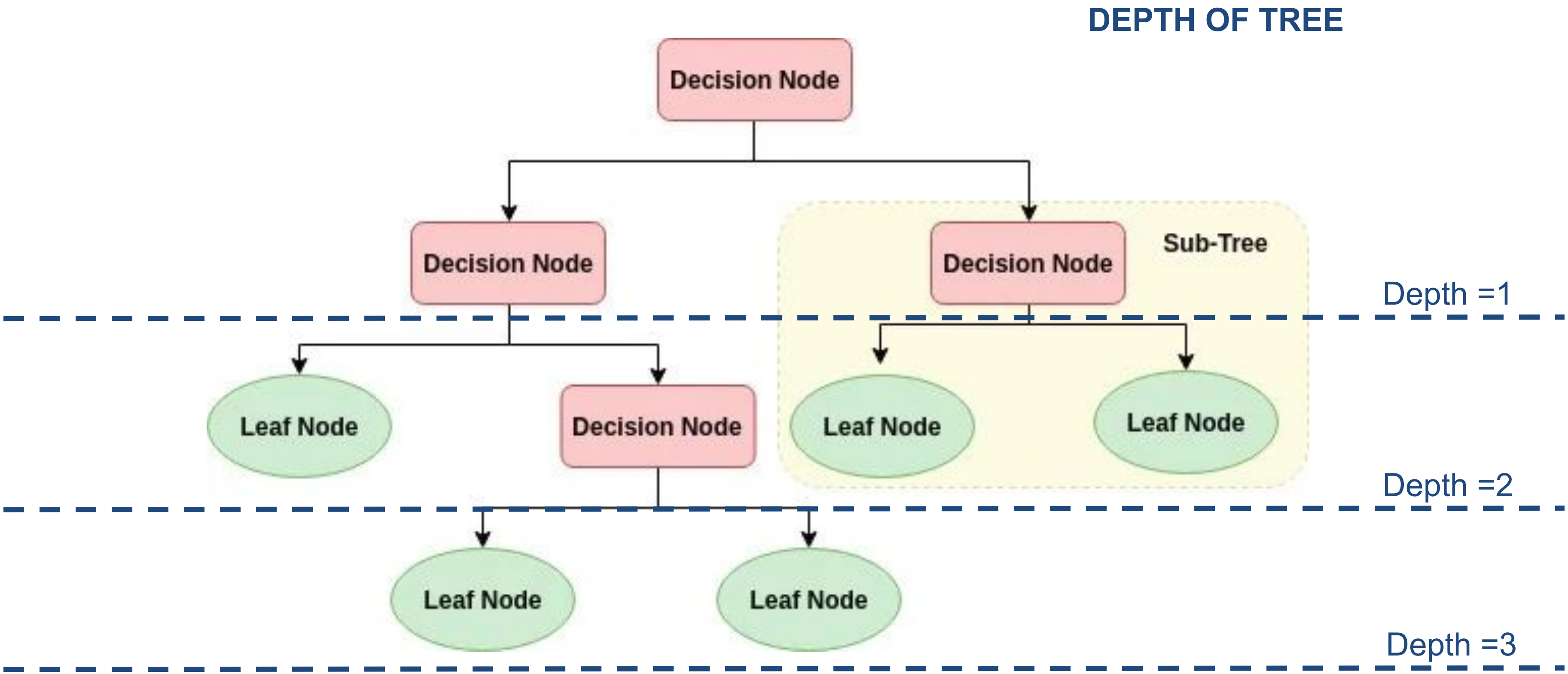


Further information about gini index: learnbymarketing.com/481/decision-tree-flavors-gini-info-gain

Decision Tree Algorithms



Decision Tree Algorithms



Random Forest

RANDOM FOREST:

The **random forest** is a model made up of many decision trees. Rather than just simply averaging the prediction of trees (which we could call a “forest”), this model uses **two key concepts** that gives it the name *random*:

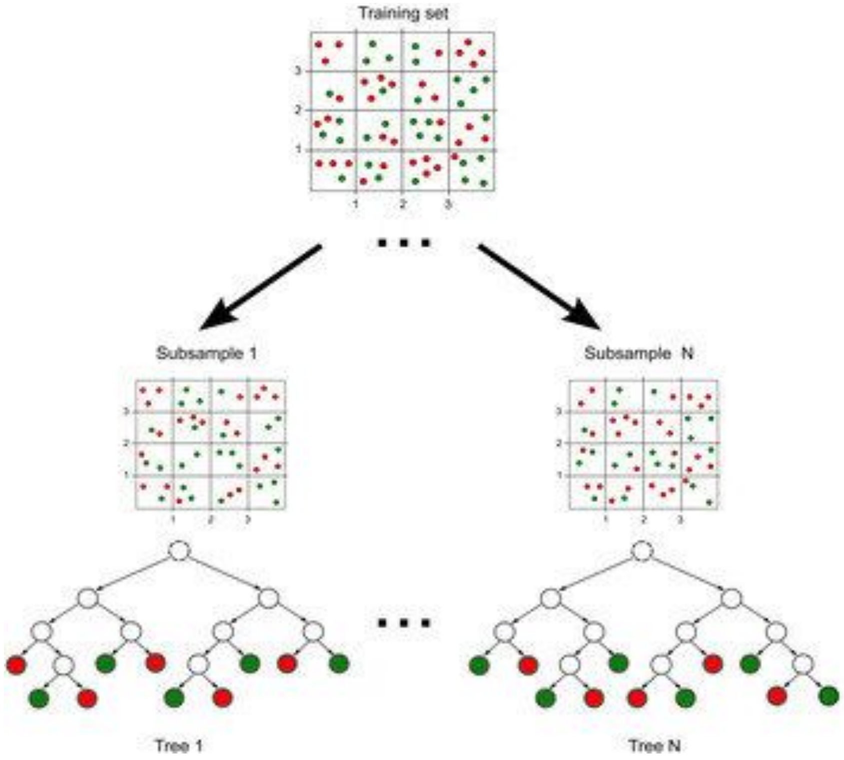
1. **Random sampling of training data** points when building trees

2. **Random subsets of features** considered when splitting nodes

Lectures: berkeley.edu - Random forest - Key Concepts

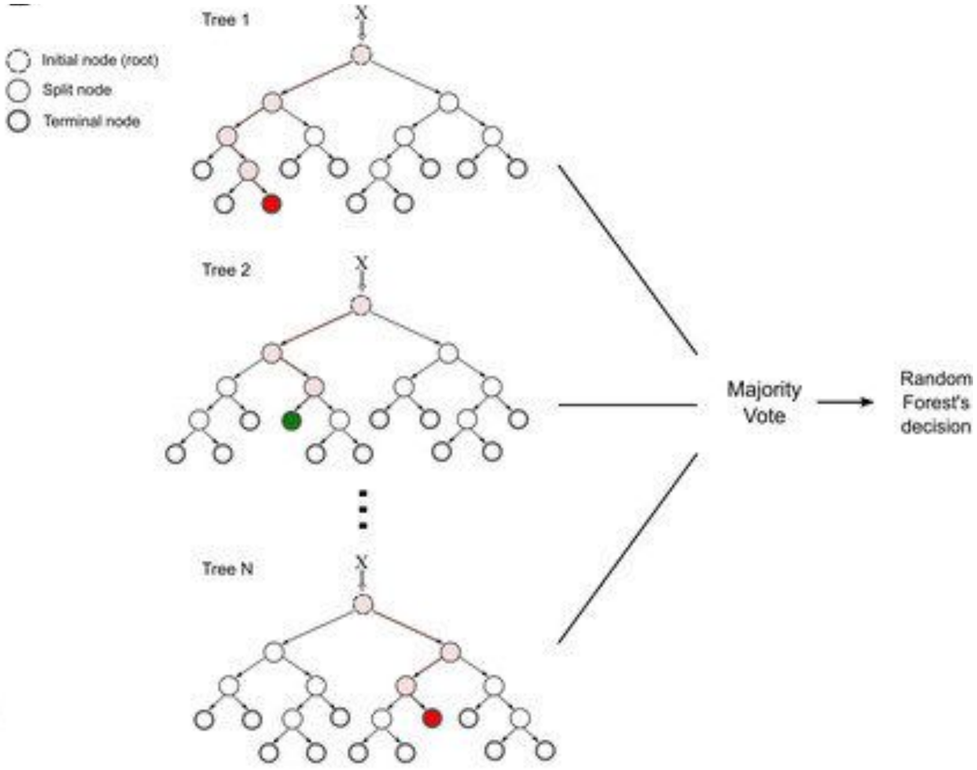
Random Forest: Random sampling of training data

Each tree in a random forest learns from a random sample of the data points.



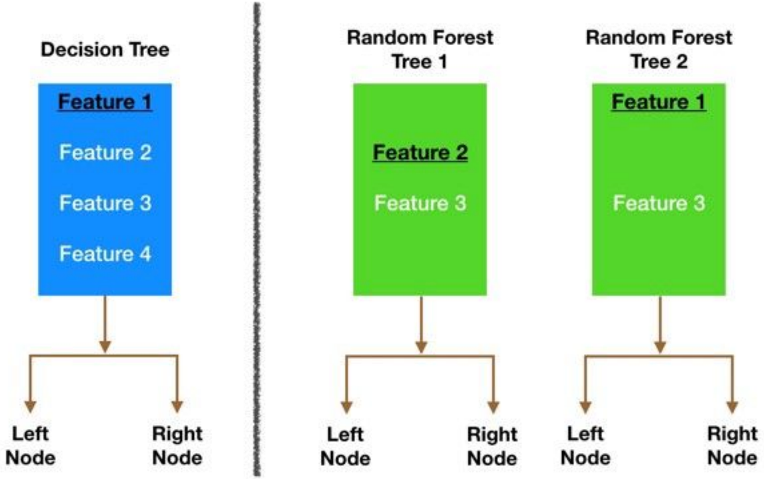
Random Forest: Random sampling of training data

At test time, predictions are made by averaging the predictions of each decision tree.



Random Forest: Random subsets of features

Only a subset of all the features are considered for splitting each node in each decision tree.
Generally this is set to sqrt(n_features) for classification meaning that if there are 16 features, at each node in each tree, only 4 random features will be considered for splitting the node.



Lectures: [Does random forest select a subset of features for every tree or every node?]

Random Forest - Intuition

Saturdays.AI
Donostia

Random Forest is the result of multiple trees that do **NOT** correlate with each other.

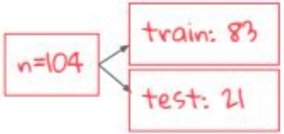
Bootstrapping

Bootstrapping

Instead of only using one holdout, we repeatedly construct different holdouts from the dataset.

Bootstrapping is a ***resampling method*** that takes random samples with replacement from whole dataset.

Example of a single holdout split. Bootstrapping repeats this many, many times. We set the number of holdouts as a hyperparameter.



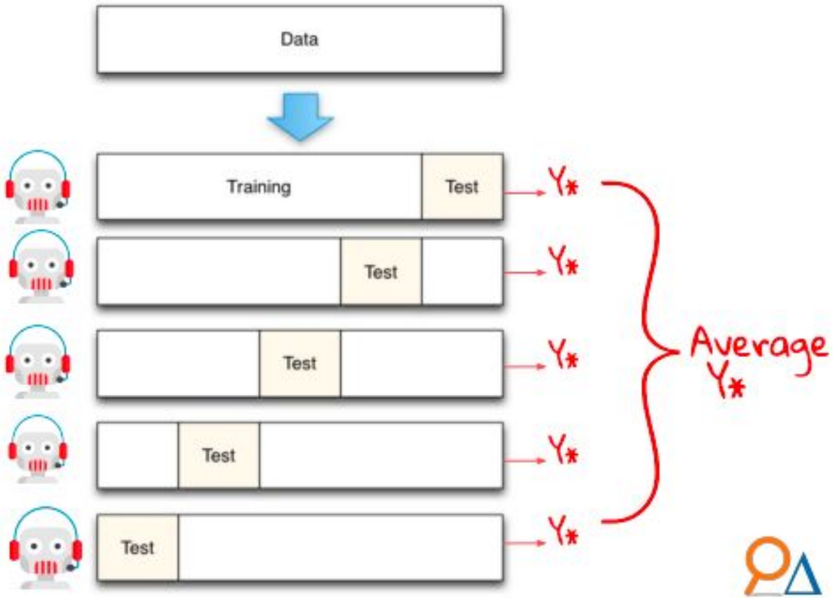
Bagging

Bagging

Bagging improves upon a single holdout by taking the average predicted Y* of boosted random samples.

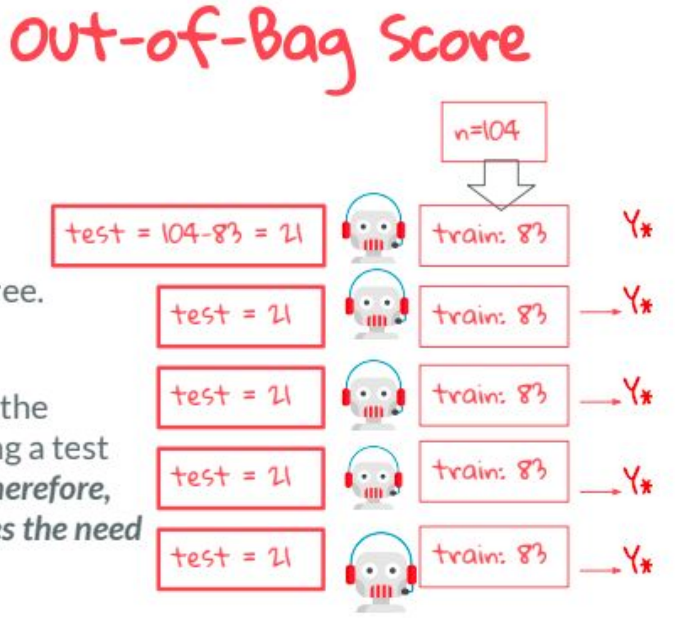
We train multiple models on random subsets of the datasets and average the predictions.

By averaging the predictions, any chance of **unrepresentative training sets** is reduced.



Out-Of-Bag

Bagging | Out-of-bag score



The out-of-bag score is the error rate of observations **not used** in each decision tree.

Why it matters:
There is empirical evidence to show that the out-of-bag estimate is as accurate as using a test set of the same size as the training set. ***Therefore, using the out-of-bag error estimate removes the need for a set-aside test set.***

Source: Breiman, 1996

Gradient Boosting Machines

Saturdays.AI
Donostia

Like Random Forests but….

Not so random….

Boosting

Saturdays.AI
Donostia

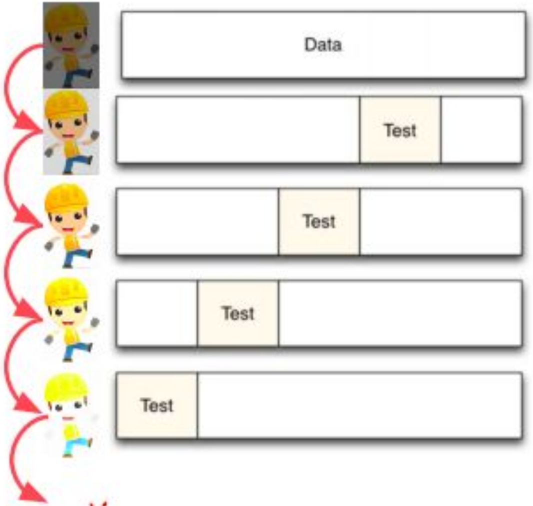
Boosting

Combining weak classifiers = one strong classifier

Boosting uses **many weak classifiers** to make a single strong classifier. A weak classifier is defined as those whose error rates is only slightly better than random guessing.

Boosting sequentially applies weak classification algorithms to repeatedly **modified versions** of the data.

How is the data modified?

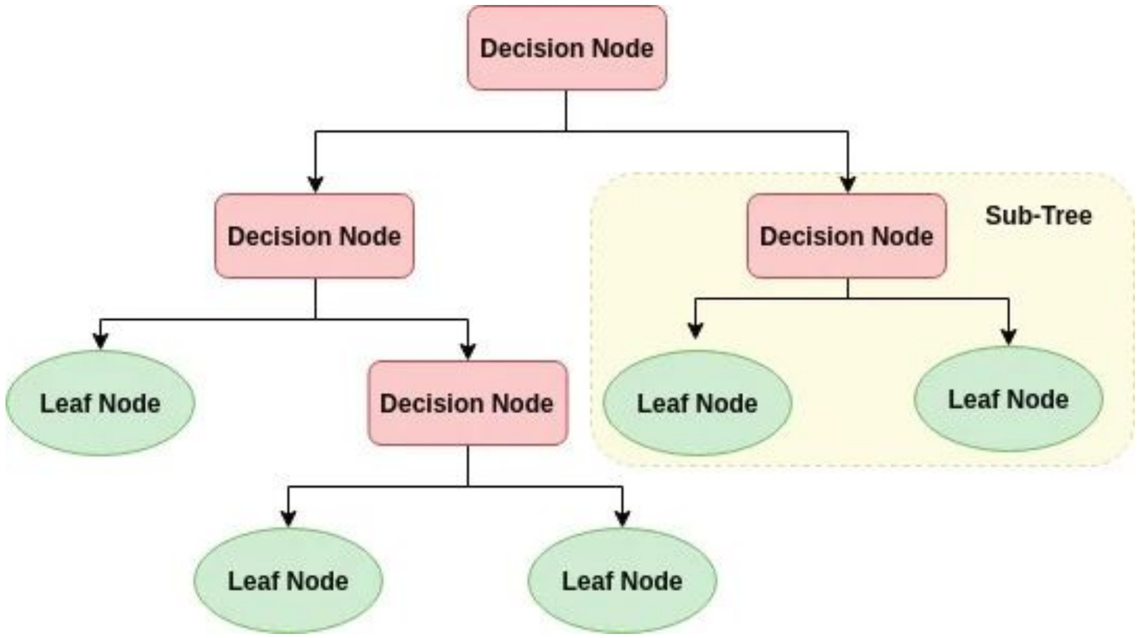


Y_*
Our Model gets "brighter"

Gradient Boosting Machines

Train 1 tree as in Random Forest

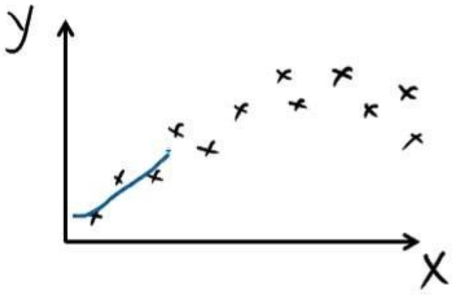
<



Gradient Boosting Machines

Calculate **cost function**

- Describes how well the current response surface $h(\boldsymbol{x})$ fits the available data (on a given data set): $J(y_i, h(\boldsymbol{x}_i))$



- Smaller values of the cost function correspond to a better fit
- Machine learning goal: construct $h(\boldsymbol{x})$ such that J is minimized
- In regression, $h(\boldsymbol{x})$ is usually directly interpretable as predicted response

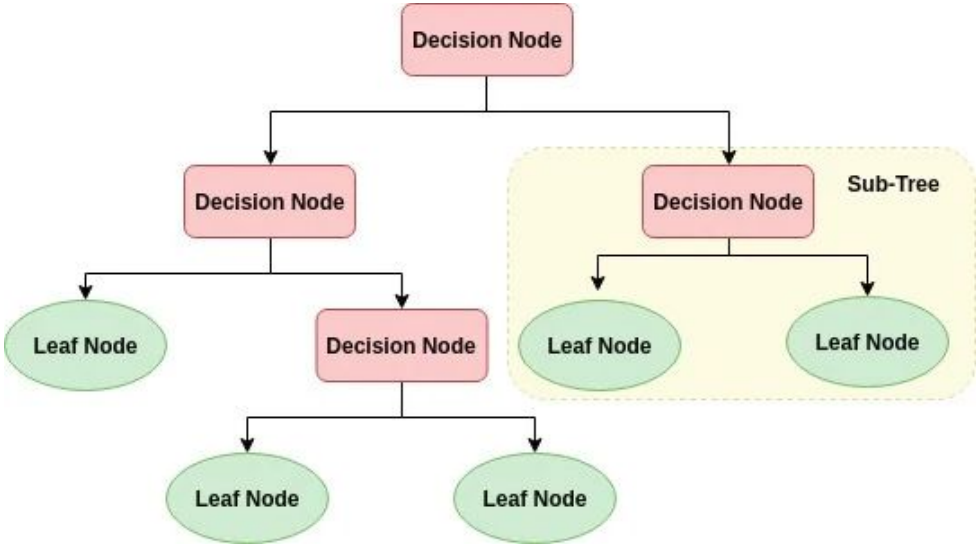
SALFORD SYSTEMS

Gradient Boosting Machines

For which data points is the tree performing worst?

Give more importance to these data points when making the next tree

Train 2nd tree considering importance

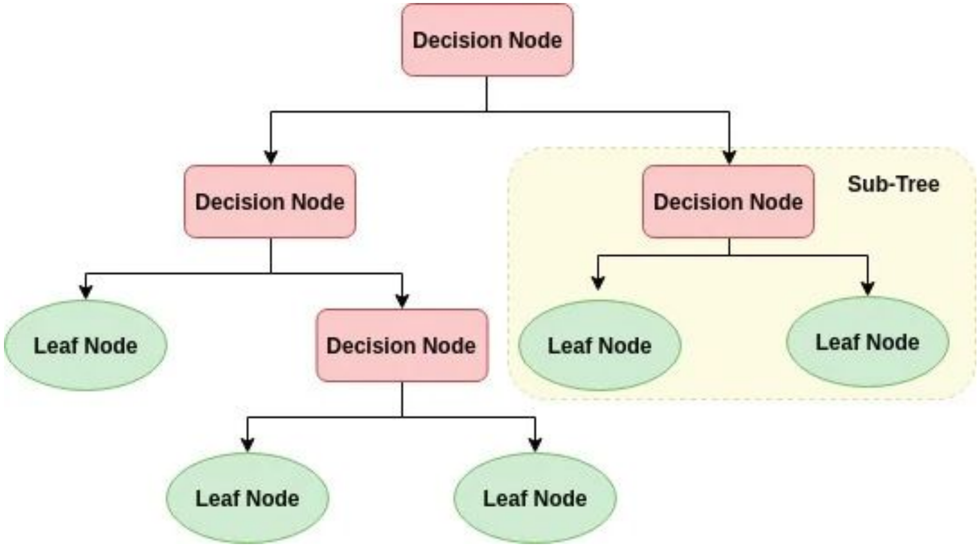


Gradient Boosting Machines

For which data points is the combination of the two trees performing worst?

Give more importance to these data points when making the next tree

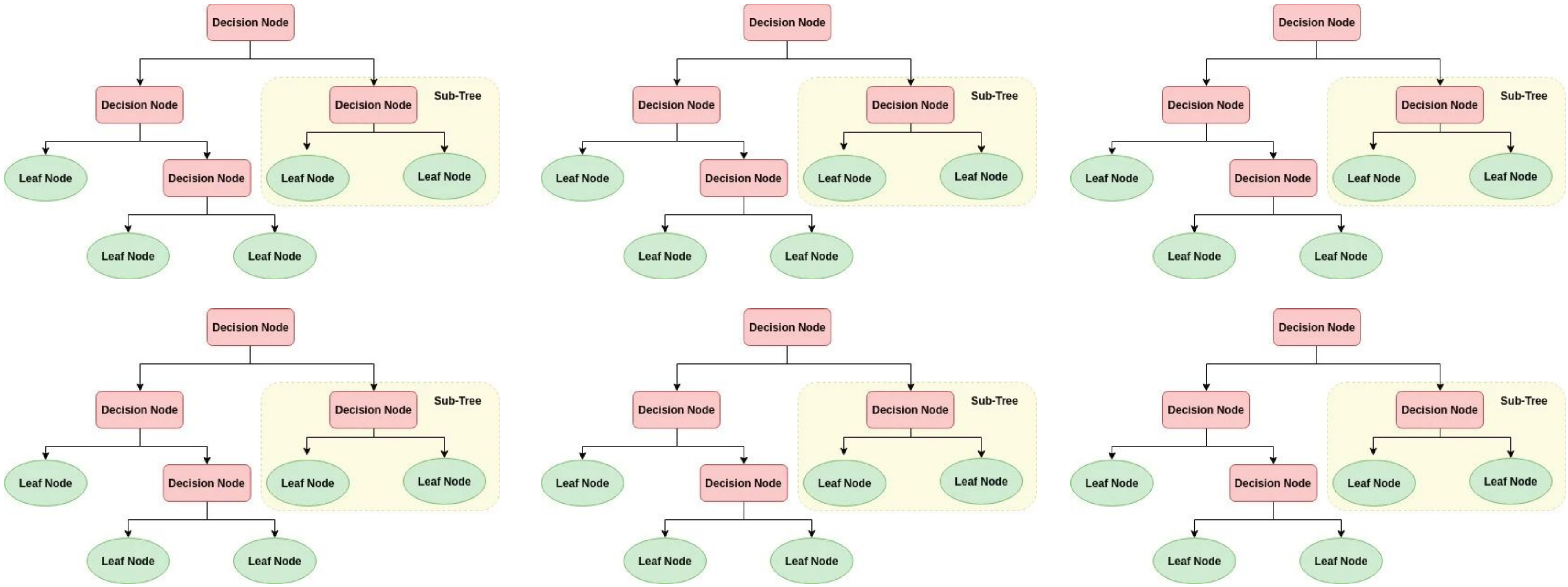
Train 3rd tree considering importance



Gradient Boosting Machines

Saturdays.AI
Donostia

Repeat with many more trees



Gradient Boosting Machines

Saturdays.AI Donostia

Available algorithms:

XGBoost

CatBoost

AdaBoost

LightGBM

....

XGBoost

Saturdays.AI Donostia

The fastest kid in town

Usually better results than plain random forest

CatBoost

Last one to join the party

Great with categorical features

Great quality without parameter tuning

Reduce time spent on parameter tuning, because CatBoost provides great results with default parameters

Categorical features support

Improve your training results with CatBoost that allows you to use non-numeric factors, instead of having to pre-process your data or spend time and effort turning it to numbers.

Fast and scalable GPU version

Train your model on a fast implementation of gradient-boosting algorithm for GPU. Use a multi-card configuration for large datasets.

Fast prediction

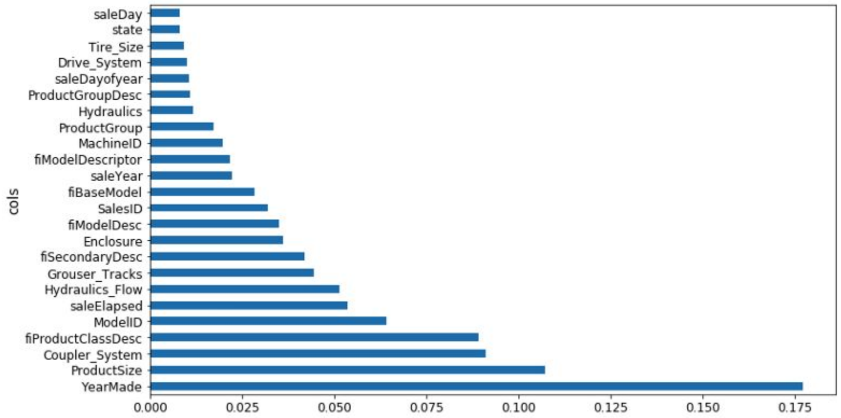
Apply your trained model quickly and efficiently even to latency-critical tasks using CatBoost's model applier

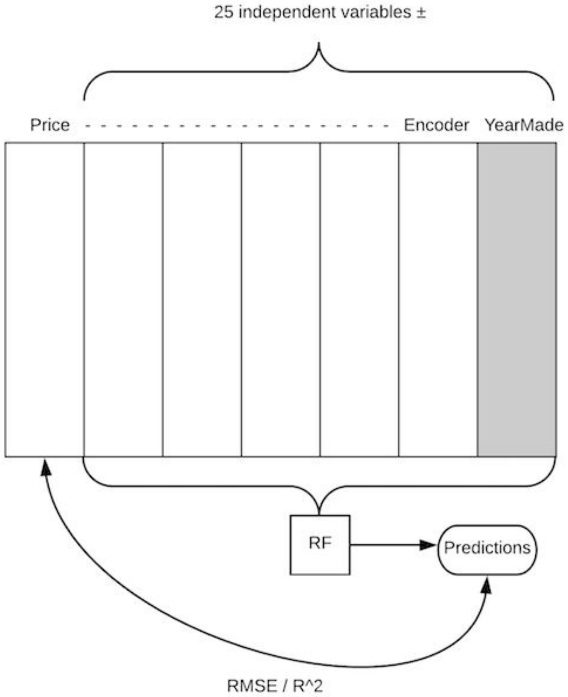
Comparison of models

4submission_xgb_5000trees_005lr_allvars_ohenom0-4_binencnom5-9_extraencallnom_nanasfeatures.pkl	3.312 KB
6submission_xgb_5000trees_005lr_allvars_featselection.pkl	3.339 KB
7submission_catboost_5000trees_005lr_allvars.pkl	448.058 KB
8submission_catboost_1000trees_01lr_allvars_categoricalfeatureset.pkl	448.058 KB
9submission_catboost_5000trees_005lr_allvars_categoricalfeatureset.pkl	1.539.165 KB

Accuracy:
4 - 0.78172
6 - 0.78129
7 - 0.78261
8 - 0.78441
9 - 0.78452

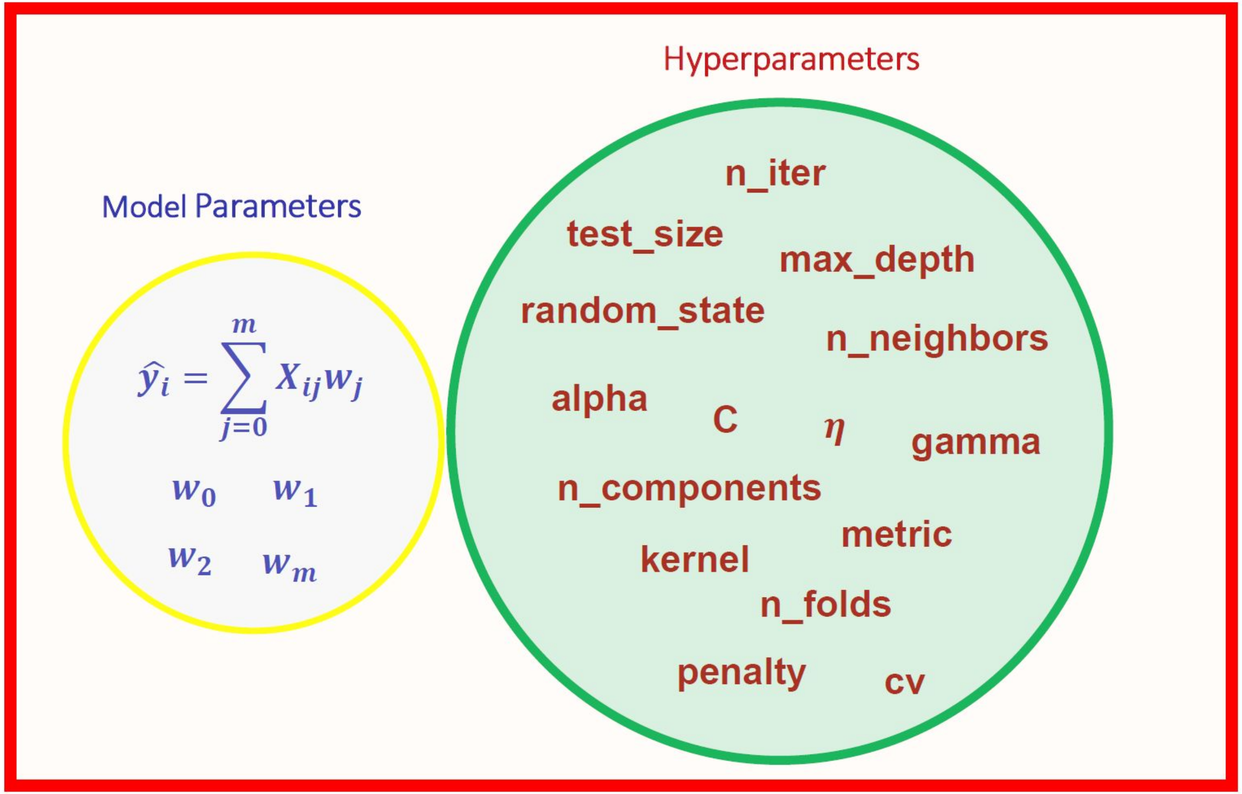
Plotting & Feature Importance





Queremos ver qué "features" hacen "puro" nuestro dataset, son las más relevantes.

Hyperparameter Tuning



More info: [1]

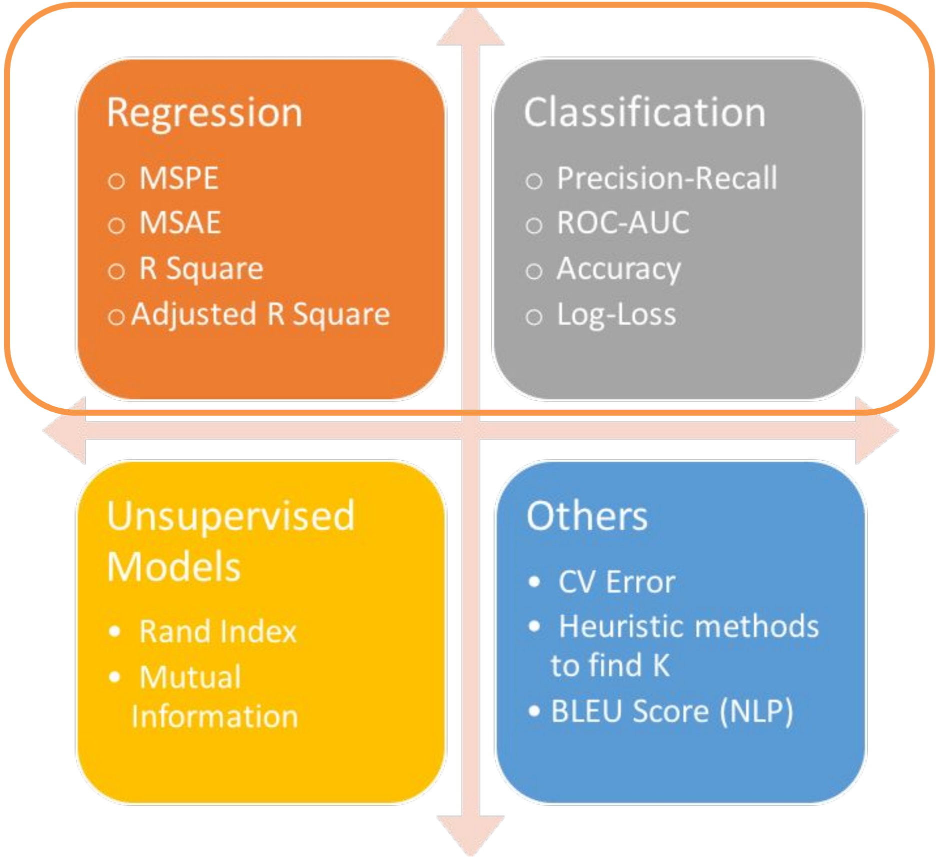
Feature Engineering

Saturdays.AI Donostia

- Hyper-parameters: the way to adjust our models
- Feature Importance: determining the most influential factors of our data, how to visualize and represent it, and solving a mistake from the last lesson in terms of sorting.
- One-Hot Encoding: how to deal with categorical data efficiently
- Remove redundant features: making our dataset more "pure" and sometimes removing the superfluous you get better results (less variance)
- Partial dependent: variables that under certain conditions may be conditioning others in a non-relational way (e.g. default values)
- Tree interpreter: taking trees one by one and dissecting per-row and per-tree predictions. Which trees and which rows provide the best efficiency/precision?
- Cross-validation: using the whole dataset as test and training at the same time (advantages and disadvantages)

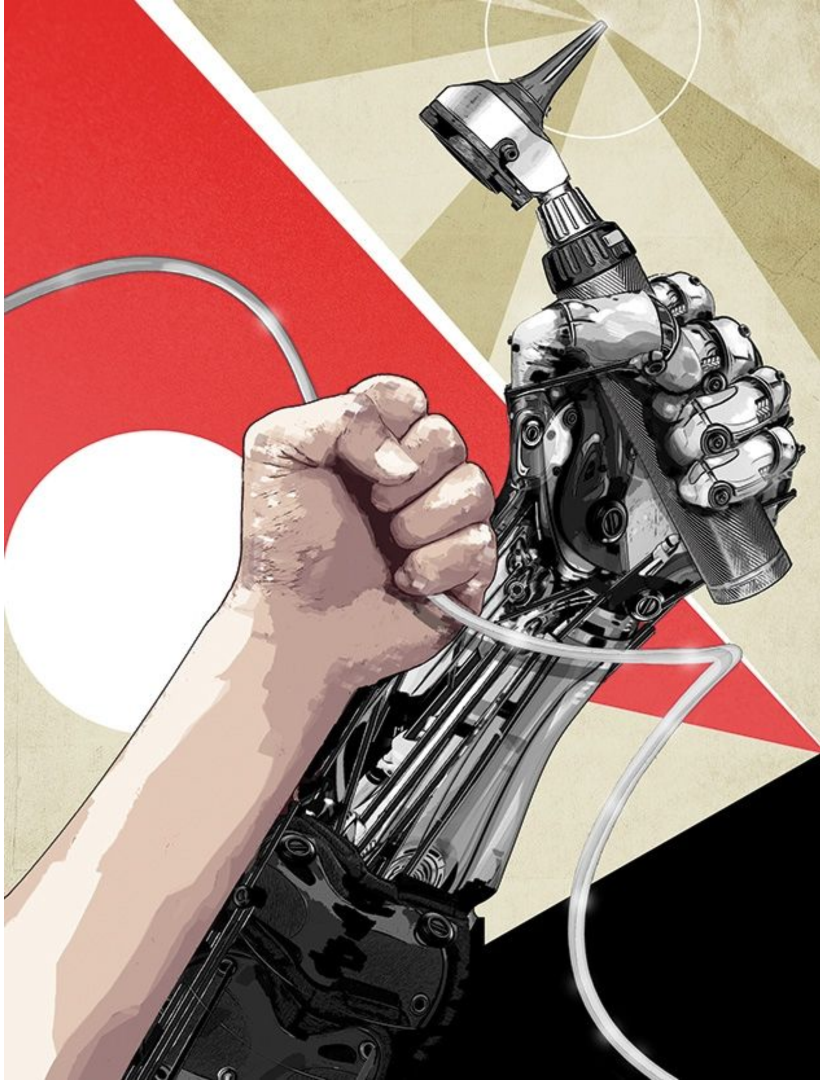
Classification vs. Regression Evaluation

Saturdays.AI
Donostia

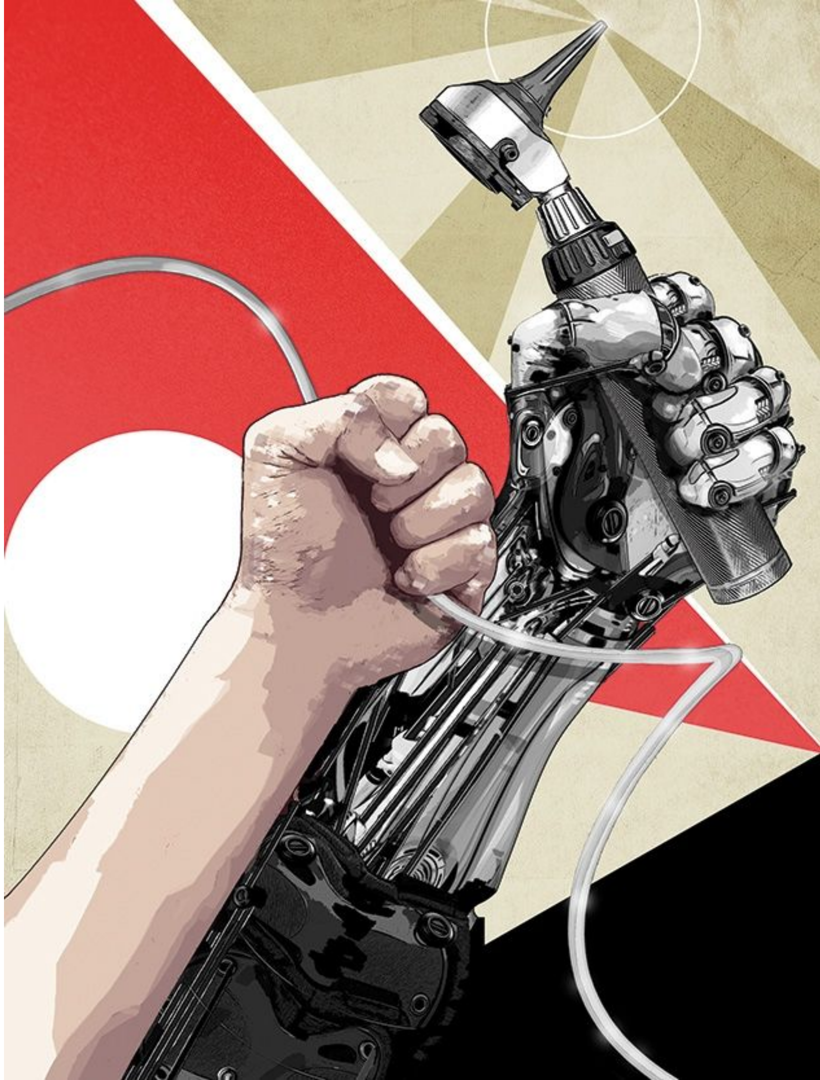


More info: [1] [2]

Practice!



Challenge!



Bibliografía

Saturdays.AI
Donostia

/1./ /Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow/

/2./ /Fast.AI - Introduction to Machine Learning for Coders/

/3./ /MLCourse.AI/

/4./ /DeltaAnalytics/

/5./ /The Hundred-page Machine Learning Book/

/6./ /Machine Learning for Humans (Vishal Maini)/

/7./ /Datacamp/

/8./ /DataQuest/

Partners

Saturdays.AI
Donostia

Agradecemos a nuestros partners por confiar en nosotros para facilitar la formación en IA de cara a la 4ª Revolución Industrial.

Saturdays.AI
Donostia

WELCOME!

Saturdays.AI

This model fits me
95% of the time

CAMP
TECNOLÓGICO

GLOBAL
SHAPERS
COMMUNITY
BILBAO

www.saturdays.ai

donostia.saturdays.ai

donostia@saturdays.ai